CSC196: Analyzing Data

Bayesian Probability and Inference

Jason Pacheco and Cesim Erten

Outline

- Introduction to Bayesian Probability
- Bayes Estimators
- Prediction and Updating
- Model Validation

Outline

- Introduction to Bayesian Probability
- Bayes Estimators
- Prediction and Updating
- Model Validation

What is Probability?

What does it mean that the probability of heads is ½ ?

Two schools of thought…

Frequentist Perspective
Proportion of successes (heads) in repeated trials (coin tosses)

Bayesian Perspective
Belief of outcomes based on assumptions about nature and the physics of coin flips

Neither is better/worse, but we can compare interpretations…

Frequentist & Bayesian Modeling

We will use the following notation throughout:

θ - Unknown (e.g. coin bias)

y - Data

Frequentist

$$p(y; \theta)$$

- θ is a non-random unknown parameter
- $p(y; \theta)$ is the *sampling / data generating distribution*

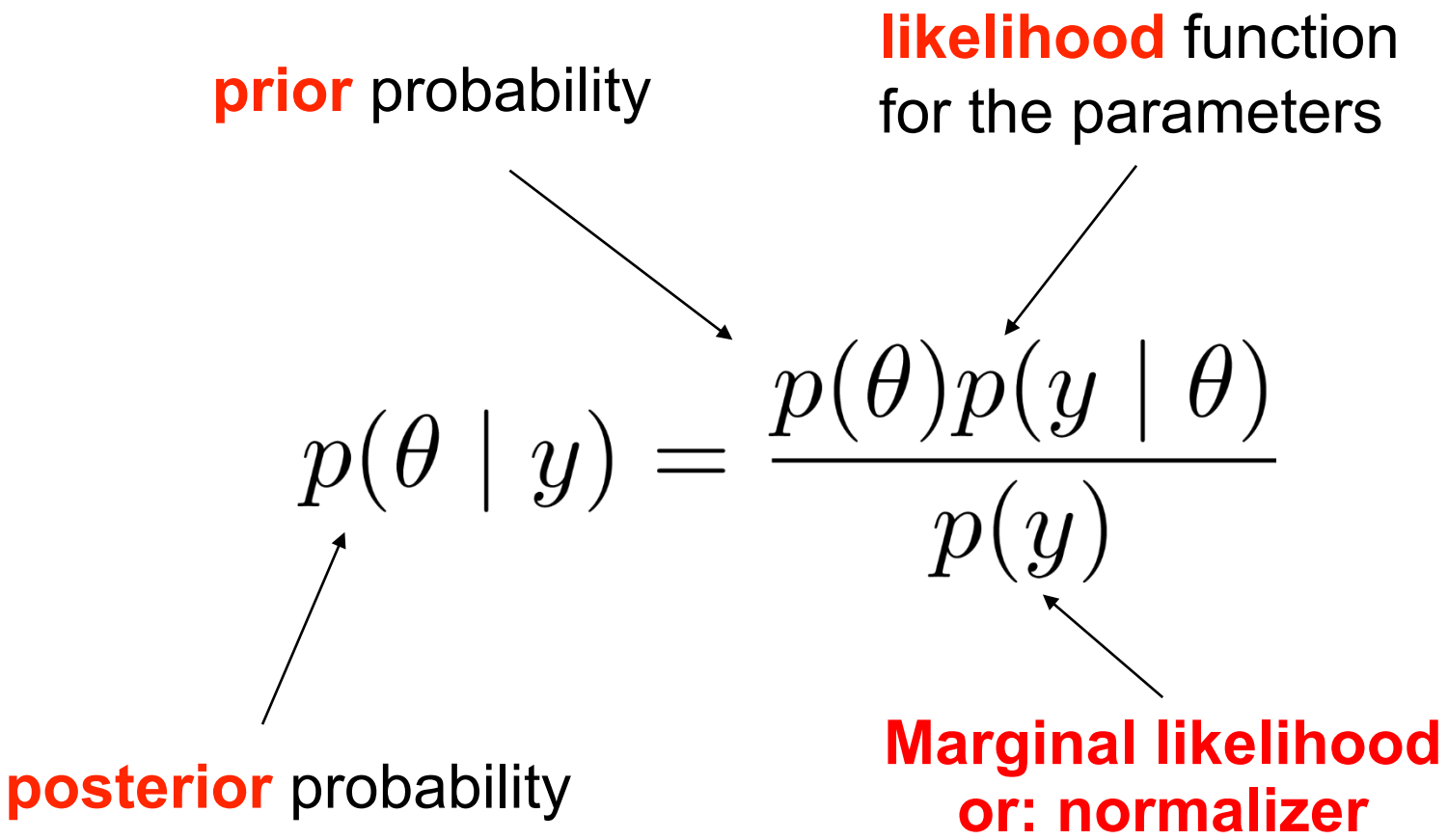
Bayesian

Prior Belief → $p(\theta)p(y \mid \theta)$ ← Likelihood

- θ is a random variable (latent)
- Requires specifying $p(\theta)$ the prior belief

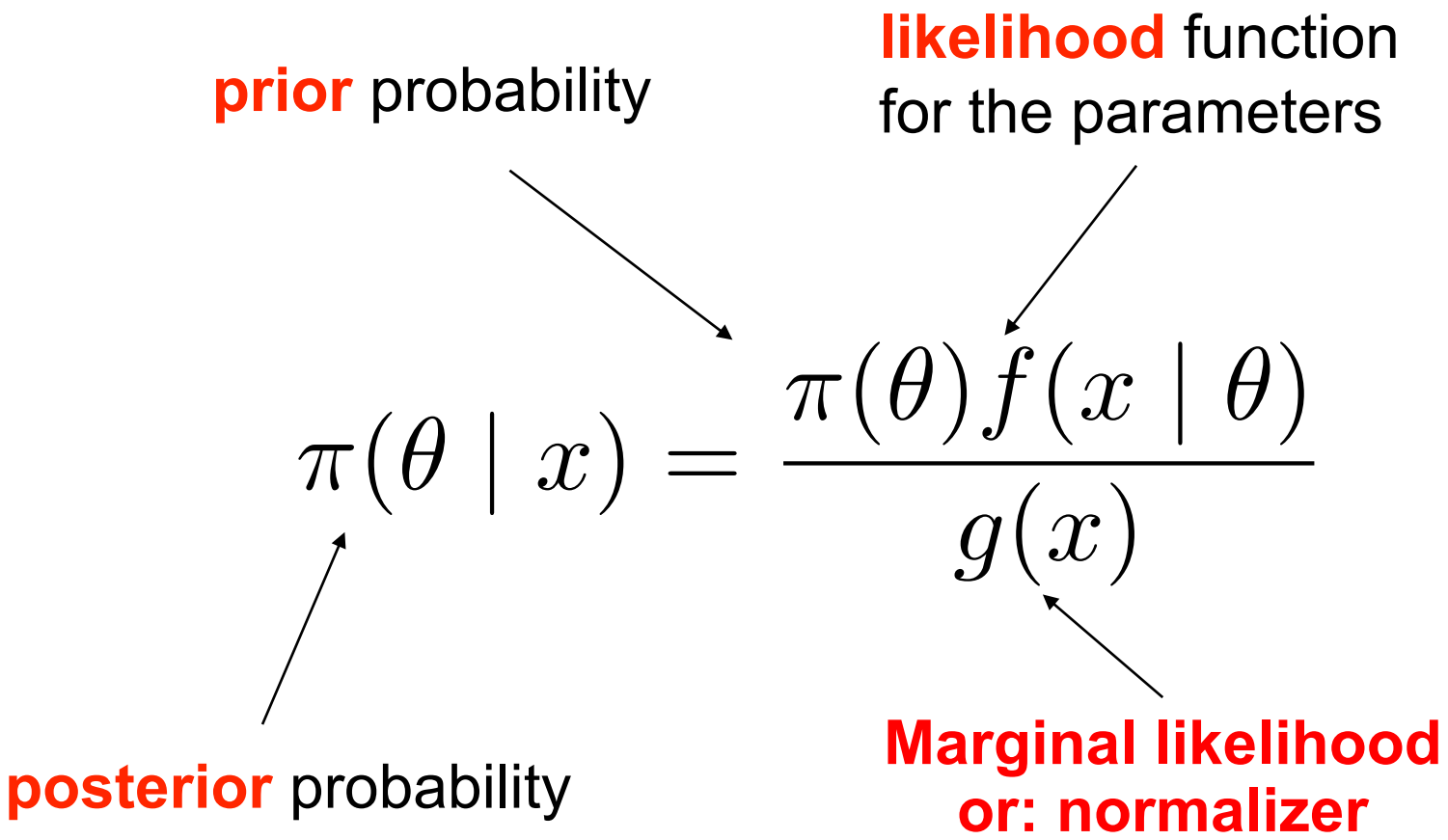
Bayes' Rule

Posterior represents all uncertainty <u>after</u> observing data…



Bayes' Rule : The Book's Notation

Posterior represents all uncertainty <u>after</u> observing data…



Priors in AI / ML / Data Science

- Priors are often used as *regularizers* (promote smoothing)
  - Reduces overfitting as random noise is not smooth
  - Often regularizers can be of simple form, even conjugate

- Priors often house sophisticated domain knowledge
  - Possibly from earlier encounters with data
  - Possibly problem constraints (e.g. θ must be nonnegative)
  - World knowledge is complex, so good priors are often complex and **not conjugate**

Bayesian Inference Example

About 29% of American adults have high blood pressure (BP). Home test has 30% false positive rate and no false negative error.

A recent home test states that you have high BP. Should you start medication?

An Assessment of the Accuracy of Home Blood Pressure Monitors When Used in Device Owners

Jennifer S. Ringrose,[1] Gina Polley,[1] Donna McLean,[2–4] Ann Thompson,[1,5] Fraulein Morales,[1] and Raj Padwal[1,4,6]

Bayesian Inference Example

About 29% of American adults have high blood pressure (BP). Home test has 30% false positive rate and no false negative error.

- Latent quantity of interest is hypertension: $\theta \in \{true, false\}$
- Measurement of hypertension: $y \in \{true, false\}$
- Prior: $p(\theta = true) = 0.29$
- Likelihood: $p(y = true \mid \theta = false) = 0.30$

$$p(y = true \mid \theta = true) = 1.00$$

Bayesian Inference Example

About 29% of American adults have high blood pressure (BP). Home test has 30% false positive rate and no false negative error.

Suppose we get a positive measurement, then posterior is:

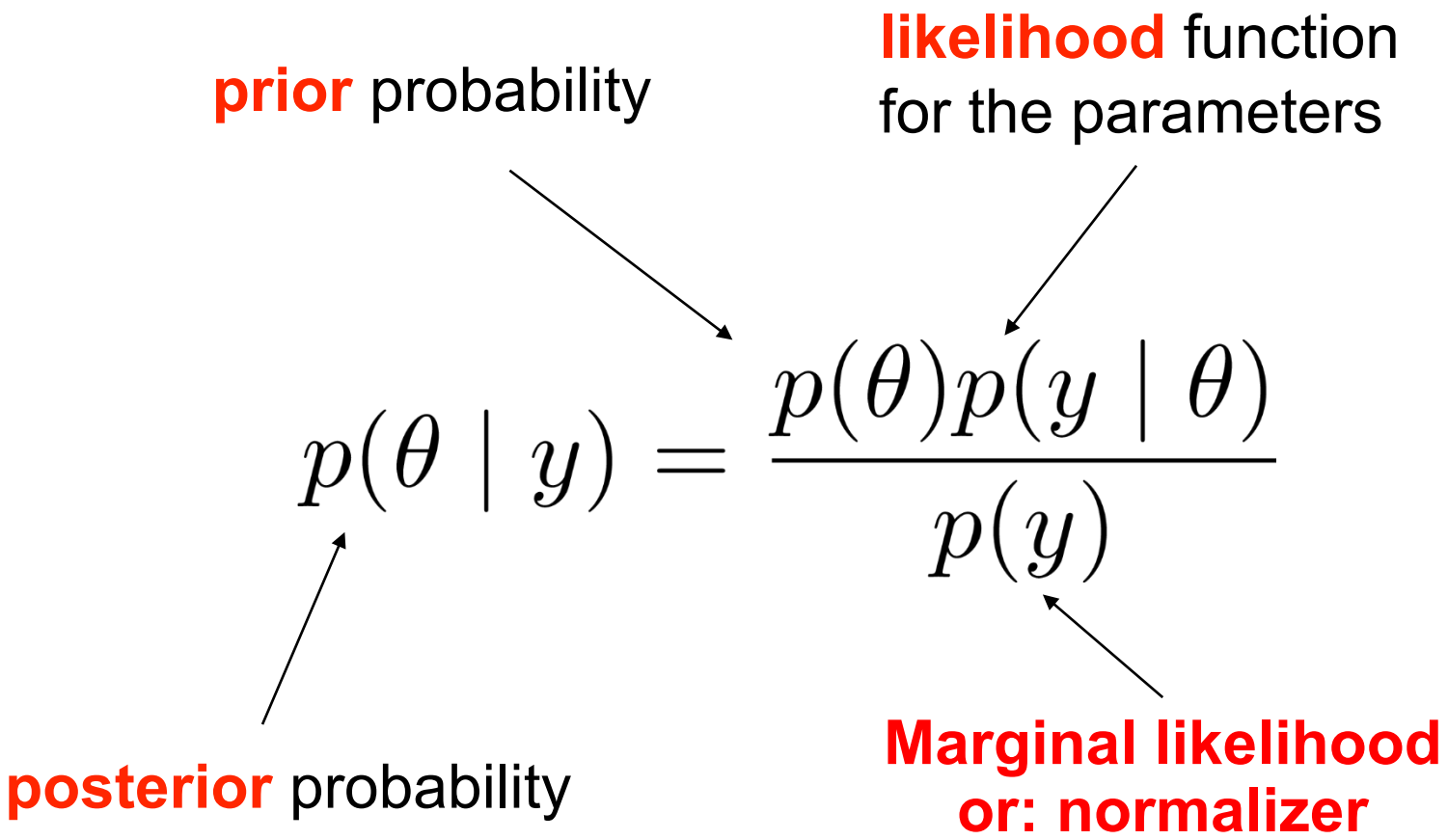
$$p(\theta = true \mid y = true) = \frac{p(\theta = true)p(y = true \mid \theta = true)}{p(y = true)}$$

$$= \frac{0.29 * 1.00}{0.29 * 1.00 + 0.71 * 0.30} \approx 0.58$$

What conclusions can be drawn from this calculation?

Bayes' Rule

Posterior represents all uncertainty <u>after</u> observing data…



Bayes' Rule : Marginal Likelihood

$$p(\theta \mid y) = \frac{p(\theta)p(y \mid \theta)}{p(y)} \propto p(\theta)p(y \mid \theta)$$

Often hard to calculate

Often know this (the model)

Marginal likelihood integrates (marginalizes) over unknown θ :

$$p(y) = \int p(\theta)p(y \mid \theta)\, d\theta$$

Marginal likelihood is less problematic in discrete models (not always)

This integral often lacks a closed form and cannot be computed…

Aside : Proportionality

Recall PMF / PDF must sum / integrate to 1,

PMF

$$\sum_x p(x) = 1$$

PDF

$$\int p(x)\,dx = 1$$

May only know distribution constant that does not depend on RV x,

$$\int \widetilde{p}(x)\,dx = \mathcal{Z} \qquad \text{so} \qquad p(x) \propto \widetilde{p}(x)$$

Properly normalized distribution by dividing our <u>normalization constant:</u>

$$\int p(x)\,dx = \int \frac{1}{\mathcal{Z}}\widetilde{p}(x)\,dx = \frac{1}{\int \widetilde{p}(x)\,dx}\int \widetilde{p}(x)\,dx = 1$$

Aside : Proportionality

Example Let X be a Bernoulli RV (coinflip) with probabilities *proportional to:*

$$\widetilde{p}(X=0) = 0.5 \qquad \widetilde{p}(X=1) = 1.5$$

Greater than 1, but It is an *unnormalized* probability

Compute normalization constant,

$$\mathcal{Z} = \widetilde{p}(X=0) + \widetilde{p}(X=1) = 2.0$$

Normalize probability distribution,

$$p(X) = \frac{1}{\mathcal{Z}}\widetilde{p}(X) = \begin{pmatrix} 1/4 \\ 3/4 \end{pmatrix}$$

Sums to 1

Outline

- Introduction to Bayesian Probability
- Bayes Estimators
- Prediction and Updating
- Model Validation

Minimum Mean Squared Error (MMSE)

Posterior mean minimizes squared error,

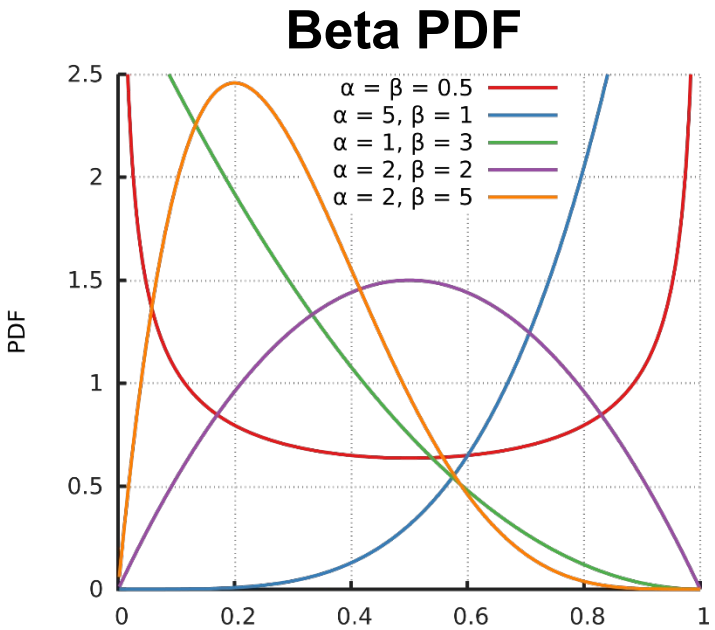
$$\hat{\theta}^{\mathrm{MMSE}} = \arg\min \mathbb{E}[(\hat{\theta} - \theta)^2 \mid y] = E[\theta \mid y]$$

- Minimizes error conditioned on observed data
- MMSE is an **unbiased estimator**
- MMSE is **asymptotically unbiased** and **asymptotically normal**,

$$\sqrt{N}(\hat{\theta}^{\mathrm{MMSE}} - \theta) \to \mathcal{N}(0, \sigma^2)$$

Example: Beta-Bernoulli MMSE

Let $X_1, \ldots, X_N \sim \text{Bernoulli}(\theta)$ and $\theta \sim \text{Beta}(\alpha, \beta)$.



- Beta is a distribution on probabilities $\theta \in [0, 1]$
- *Shape* parameters α and β with mean,

$$\mathbf{E}[\theta] = \frac{\alpha}{\alpha+\beta}$$

- Beta-Bernoulli has Beta posterior distribution,

$$p(\theta \mid X_1^N) = \text{Beta}(\alpha + \textbf{number of heads},\ \beta + \textbf{number of tails})$$

MMSE given by posterior mean,

Prior belief (pseudo-heads)

$$\hat{\theta}^{\text{MMSE}} = \frac{\alpha + \textbf{number of heads}}{\alpha + \beta + N}$$

Q What happens to MMSE when we have limited data?

Q What happens to MMSE when we have a lot of data?

Bayes Estimators

Minimizes expected loss function,

$$\hat{\theta} = \arg\min_{\hat{\theta}} \mathbf{E}\left[L(\theta, \hat{\theta}) \mid y\right]$$

Expected loss referred to as *Bayes risk.*

MMSE minimizes squared-error loss $L(\theta, \hat{\theta}) = (\theta - \hat{\theta})^2$

Minimum absolute error (MAE) is posterior *median,*

$$\arg\min \mathbb{E}[|\hat{\theta} - \theta| \mid y] = \text{median}(\theta \mid y)$$

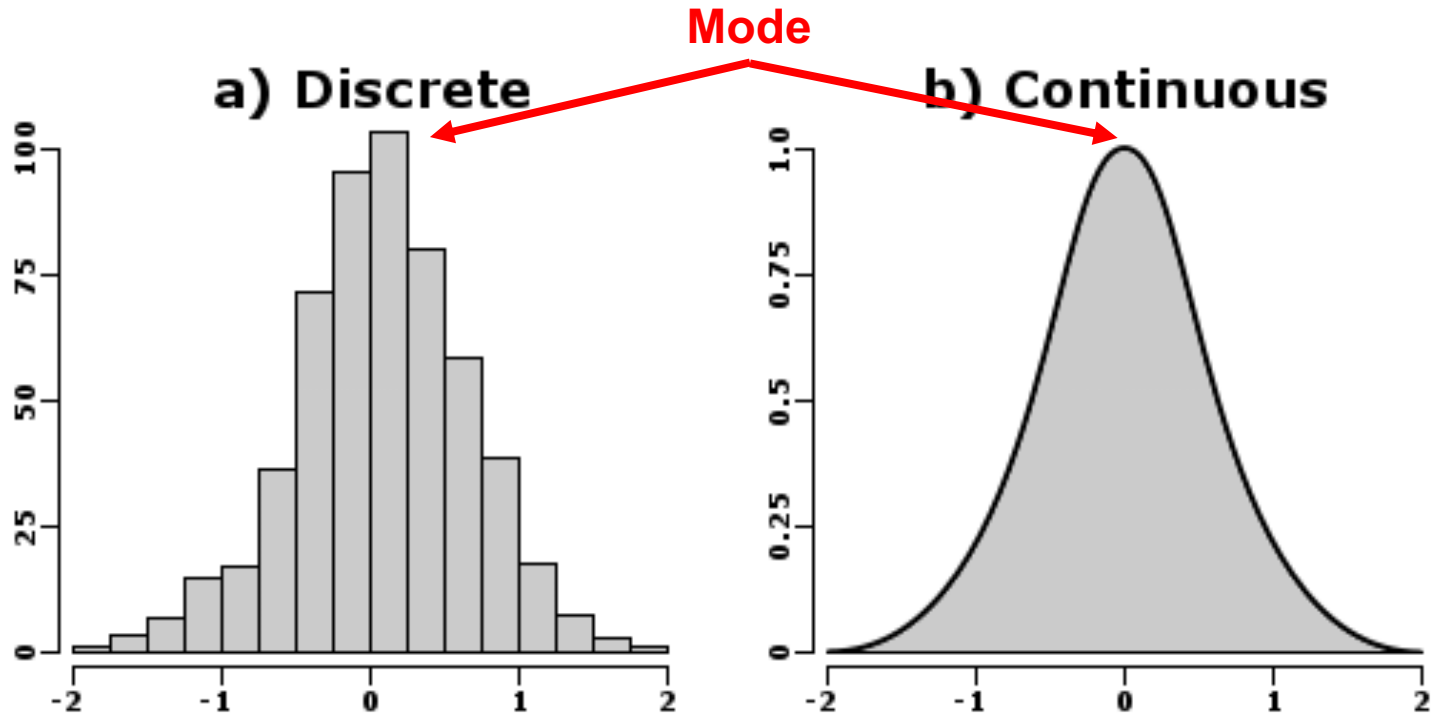
Note: Same answer for linear function: $L(\theta, \hat{\theta}) = c|\hat{\theta} - \theta|$

Maximum a Posteriori (MAP)

Very common to produce maximum probability estimates,

$$\hat{\theta}^{\mathrm{MAP}} = \arg max\ p(\theta \mid y)$$

*MAP is the **mode** (highest probability outcome) of the posterior*



Example: Beta-Bernoulli MAP

Let $X_1, \ldots, X_N \sim \text{Bernoulli}(\theta)$ and $\theta \sim \text{Beta}(\alpha, \beta)$ then posterior is,

$$p(\theta \mid X_1^N) = \text{Beta}(\alpha + \underbrace{\textbf{number of heads}}_{N_H}, \ \beta + \textbf{number of tails})$$

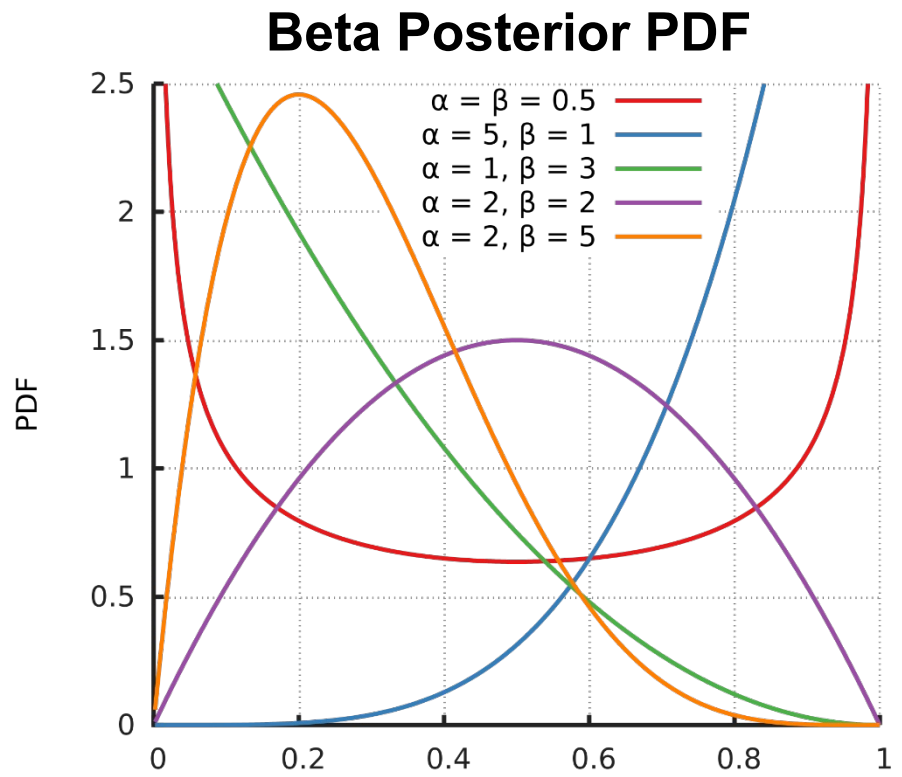
Highest probability (mode) of Beta given by,

Take derivative, set to zero, solve.

$$\hat{\theta}^{\text{MAP}} = \frac{\alpha + N_H - 1}{\alpha + \beta + N - 2}$$

Beta distribution is not always convex!

- MAP is any value for $\alpha = \beta = 1$
- Two modes (bimodal) for $\alpha, \beta < 1$



Maximum a Posteriori (MAP)

Equivalent to maximizing joint probability,

$$\arg\max_\theta p(\theta \mid y) = \arg\max_\theta \frac{p(\theta, y)}{p(y)} = \arg\max_\theta p(\theta, y)$$

Constant

For iid $y_1, \ldots, y_N$ solve in log-domain (like *maximum likelihood est.*),

$$\hat{\theta}^{\text{MAP}} = \arg\max_\theta \log p(\theta, y_1, \ldots, y_N) = \underbrace{\sum_i \log p(y_i \mid \theta)}_{\text{Log-Likelihood (how well it fits data)}} + \underbrace{\log p(\theta)}_{\text{Log-Prior (how well it agrees with prior)}}$$

Intuition *MAP is like MLE but with a "penalty" term (log-prior)*

Outline

- Introduction to Bayesian Probability
- Bayes Estimators
- Prediction and Updating
- Model Validation

Prediction

Can make predictions of unobserved $\tilde{y}$ before seeing any data,

$$p(\tilde{y}) = \sum_k p(\theta = k) p(\tilde{y} \mid \theta = k)$$

Similar calculation to marginal likelihood

*This is the **prior predictive** distribution*

For continuous parameters sum turns into integral,

$$p(\tilde{y}) = \int p(\theta) p(\tilde{y} \mid \theta) \, d\theta$$

*This is a prediction based on **no observed data***

Prediction

When we observe y we can predict future observations $\tilde{y}$,

$$p(\tilde{y} \mid y) = \sum_k \underbrace{p(\theta = k \mid y)} p(\tilde{y} \mid \theta = k)$$

This is now the posterior

*This is the **posterior predictive** distribution*

Again, for continuous parameters sum turns into integral,

$$p(\tilde{y} \mid y) = \int p(\theta \mid y) p(\tilde{y} \mid \theta) \, d\theta$$

Prediction Example

About 29% of American adults have high blood pressure (BP). Home test has 30% false positive rate and no false negative error.

What is the likelihood of *another* positive measurement?

$$p(\tilde{y} = true \mid y = true) = \sum_{\theta \in \{true, false\}} p(\theta \mid y = true) p(\tilde{y} = true \mid \theta)$$

$$= 0.42 * 0.30 + 0.58 * 1.00 \approx 0.71$$

What conclusions can be drawn from this calculation?

Bayesian Updating

Suppose we plan to take another test…

Question What is our belief about blood pressure status *before* the second test?

(a) Posterior: $p(\theta = true \mid y_1 = true)$

(b) Likelihood: $p(y_1 = true \mid \theta = true)$

(c) Marginal Likelihood: $p(y_1 = true)$

Bayesian Updating

Suppose we plan to take another test…

Question What is the probability that we get *true* on the second test if we have high blood pressure?

(a) Posterior: $p(\theta = true \mid y_1 = true, y_2 = true)$

(b) Likelihood: $p(y_2 = true \mid \theta = true)$

(c) Marginal Likelihood: $p(y_2 = true)$

Why not: $p(y_2 = true \mid \theta = true, y_1 = true)$

Bayesian Updating

Suppose we plan to take another test…

Question What is the probability that we get *true* on the second test if we have high blood pressure?

(a) Posterior: $p(\theta = true \mid y_1 = true, y_2 = true)$

(b) Likelihood: $p(y_2 = true \mid \theta = true)$

(c) Marginal Likelihood: $p(y_2 = true)$

Because $y_1 \perp y_2 \mid \theta$ *so* $p(y_2 \mid \theta, y_1) = p(y_2 \mid \theta)$

Outline

- Introduction to Bayesian Probability
- Bayes Estimators
- Prediction and Updating
- Model Validation

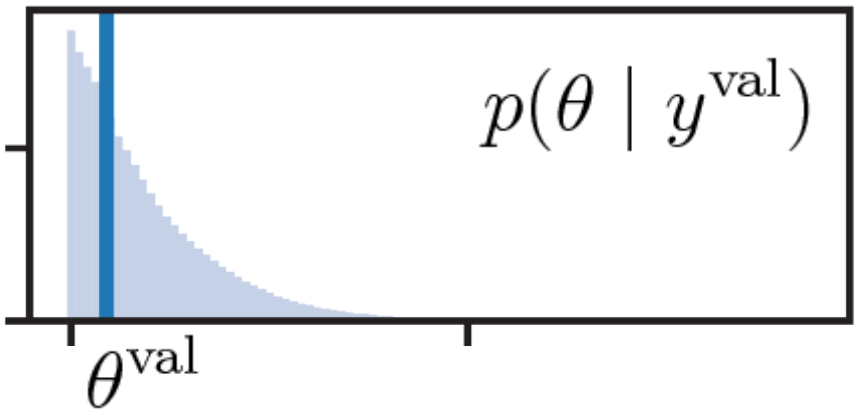
Model Validation

How do we know if the model $p(\theta, y)$ is good?

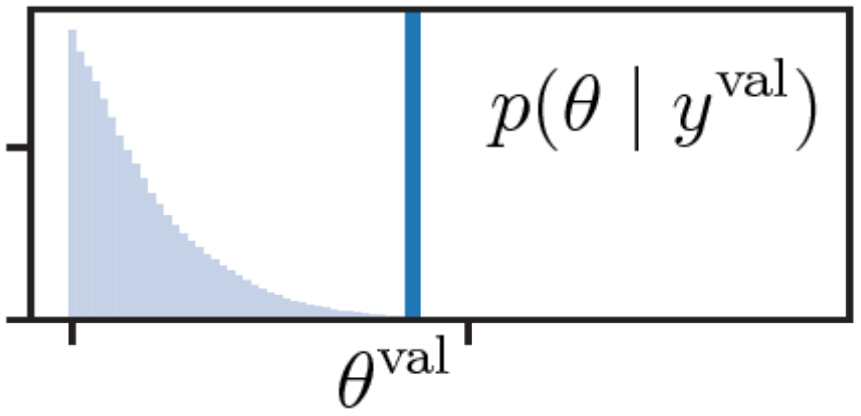
Supervised Learning

Validation set $\{(\theta^{\text{val}}, y^{\text{val}})\}$ consists of known θ^{val}. Are true values typically preferred under the posterior?

Good (maybe lucky)



Not Good (maybe unlucky)



Repeat trials over validation set for more certainty

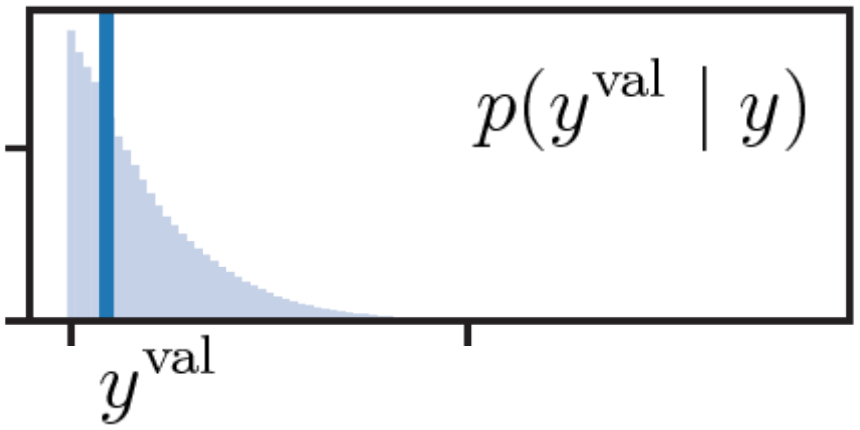
Model Validation

How do we know if the model $p(\theta, y)$ is good?

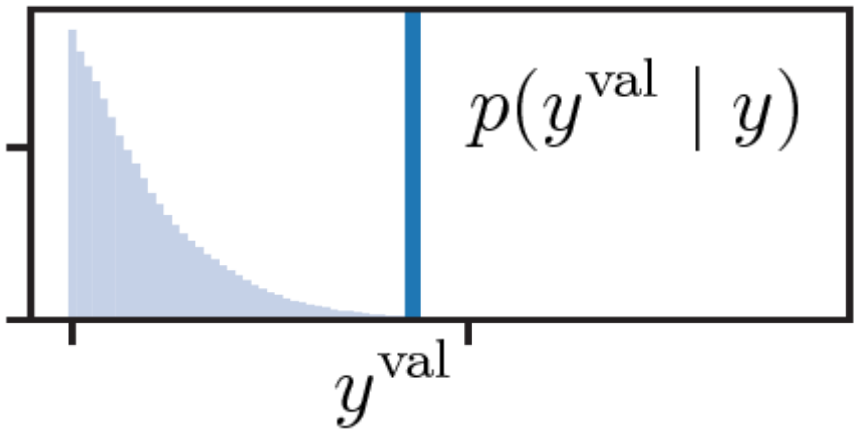
Unsupervised Learning

Validation set $\{y^{\text{val}}\}$ only contains observable data. Check validation data against posterior-predictive distribution.

Good (maybe lucky)



Not Good (maybe unlucky)



Repeat trials over validation set for more certainty

Likelihood and Odds Ratios

Which parameter value θ_1 or θ_2 is more likely to have generated the observed data y?

The **posterior odds ratio** is:

$$\frac{p(\theta_1 \mid y)}{p(\theta_2 \mid y)} = \frac{p(\theta_1)}{p(\theta_2)} \frac{p(y \mid \theta_1)}{p(y \mid \theta_2)} \frac{p(y)}{p(y)}$$

Prior Odds Ratio

Likelihood Ratio

Observe: the marginal likelihood $p(y)$ cancels!

Posterior Summarization

Ideally we would report the full posterior distribution as the result of inference…but this is not always possible

Summary of Posterior Location:

Point estimates: mean (MMSE), mode, median (min. absolute error)

Summary of Posterior Uncertainty:

Credible intervals / regions, posterior entropy, variance

Bayesian analysis should report uncertainty when possible

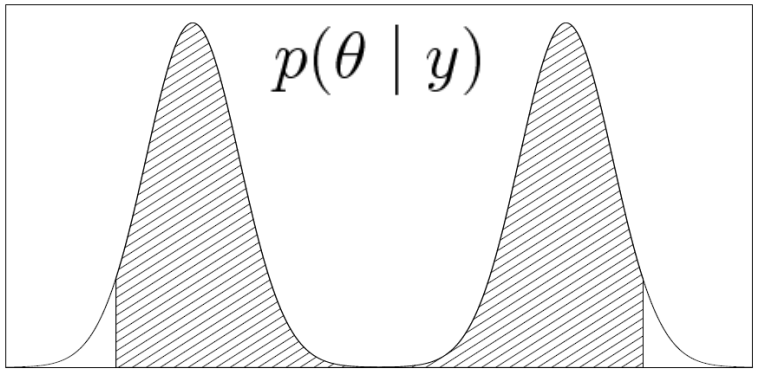
Credible Interval

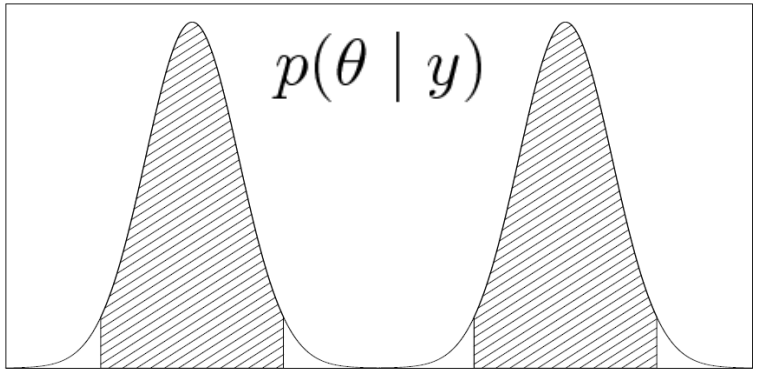
Def. For parameter $0 < \alpha < 1$ the $100(1-\alpha)\%$ credible interval $(L(y), U(y))$ satisfies,

$$p(L(y) < \theta < U(y) \mid y) = \int_{L(y)}^{U(y)} p(\theta \mid y) = 1 - \alpha$$

Interval containing fixed percentage of posterior probability density.

Note: This is <u>not unique</u> -- consider the 95% intervals below:





[Source: Gelman et al., “Bayesian Data Analysis”]

Frequentist Inference

Example: Suppose we observe the outcome of N coin flips. $y = \{y_1, \ldots, y_N\}$. What is the probability of heads θ (coin bias)?

- Coin bias θ is not random (e.g. there is some *true* value)
- Uncertainty reported as confidence interval (typically 95%)

  Correct Interpretation: On repeated trials of N coin flips θ will fall inside the confidence interval 95% of the time (in the limit)

- Inferences are valid for multiple trials, **never on single trials**

  Wrong Interpretation: For *this trial* there is a 95% chance θ falls in the confidence interval

Bayesian Inference

Posterior distribution is complete representation of uncertainty

$$p(\theta \mid y) = \frac{p(\theta)p(y \mid \theta)}{p(y)}$$

Prior Belief

Likelihood

Marginal Likelihood (more on this later)

- Must specify a prior belief $p(\theta)$ about coin bias
- Coin bias θ is a random quantity
- Interval $p(l(y) < \theta < u(y) \mid y) = 0.95$ can be reported in lieu of full posterior, and takes intuitive interpretation for a single trial

Interval Interpretation: For this experiment there is a 95% chance that θ lies in the interval

Summary

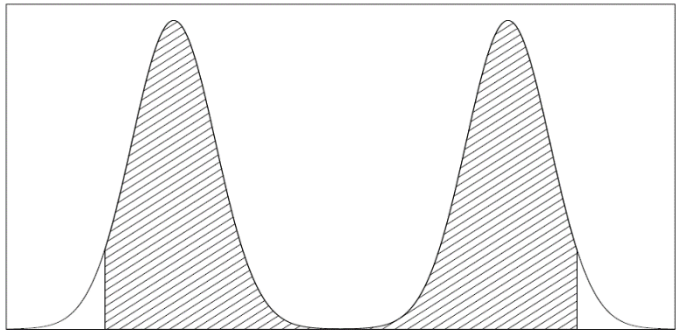
- Bayesian statistics interprets probability differently than classical stats
  - Frequentist: Probability → Long run odds in repeated trials
  - Bayesian: Probability → Belief of outcome that captures all uncertainty
- Bayesian models treat unknown parameter as random, with a prior
- Bayesian inference via the *posterior distribution* using Bayes' rule

$$p(\theta \mid y) = \frac{p(\theta)p(y \mid \theta)}{p(y)}$$

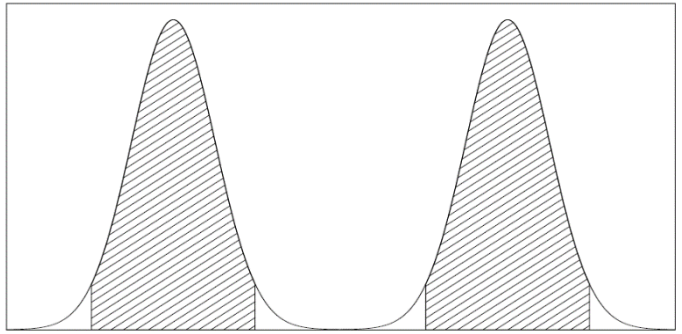
- Bayesian estimators minimize expected risk (e.g. MMSE)
- Maximum a posteriori (MAP) estimate maximizes posterior probability

Summary

- Conjugate prior-posterior pairs ensure closed-form posterior inference
- Posterior uncertainty can be characterized by credible intervals



Not necessarily unique



- Selecting models can be done via posterior odds ratio

$$\frac{p(\theta_1 \mid y)}{p(\theta_2 \mid y)} = \frac{p(\theta_1)}{p(\theta_2)} \frac{p(y \mid \theta_1)}{p(y \mid \theta_2)} \frac{p(y)}{p(y)}$$

- Parameter can be marginalized out via prior/posterior predictive dist'n